A linguistic approach towards the functional characterization of intragenic non-exon sequences

Menconi G[1], Conti V[2], Puliti A[2], Sbrana I[3], Marangoni R[4,5]

[1] Istituto Nazionale di Alta Matematica, Roma
[2] Molecular Genetics and Cytogenetics Unit, G. Gaslini Institute, Genova, Italy and Renal Child Foundation, G. Gaslini Institute, Genova, Italy
[3] Department of Biology, University of Pisa, Italy.
[4] Department of Computer Science, University of Pisa, Italy.
[5] CNR - Institute of Biophysics, Pisa, Italy

Motivation

Linguistic methods represent a very promising approach to different potential applications in bioinformatics. In particular, methods for text compression have been applied to coding/non-coding sequences classification and they are often based on dictionaries: the compression algorithm builds up a collection of words that are good to use as prefixes or suffixes of other words in the text. In a previous work, a suitable dictionary-based compression algorithm (CASToRe) has been used to predict coding sequences in prokaryotic genomes, leading to very good results, comparable with those of the most widely used machine-learning based methods. By inspecting the words stored by CASToRe in the dictionary, we discovered that the compression of coding sequences lead to the automatic storage of commonly known genetic code triplets. More intriguing was the inspection of the dictionary after the compression of non-coding sequences. Since prokaryota possess, in general, very short non-coding sequences, it is more interesting to explore them in eukaryota. The present work is focused on the analysis of dictionary words generated during the compression of some very long intragenic non-exons sequences of two couples of paralogous genes, the metabotropic Glutamate receptors 1 (GRM1) and 5 (GRM5) in the human and in the mouse.

Methods

The algorithm CASToRe belongs to the Lempel-Ziv family and selects a dictionary by exact matches and parses the input sequence T in some variable-length recurrent words. Each new parsed word is the one that can be made with the longest prefix and the longest suffix already in the dictionary. The input sequence T is parsed in pairwise distinct subwords belonging to the final dictionary relative to the sequence. We calculate the word score as follows. Each word w occurs occ(w)=1 times within the sequence T when used as a prefix or a suffix of a subsequent word. The score of w is obtained from its length and its occurrence rate. The most interesting words are long and have a high score. They are selected as the most representative of the structured motifs characteristic of the input sequence.

Results

CASToRe has been applied to the analysis of four genes, the metabotropic Glutamate receptors 1 and 5 in the human and in the mouse. Human sequences were from NCBI Build 36.1 and mouse sequences from Build 37 (UCSC Genome Bioinformatics, http://genome.ucsc.edu). We based our analysis of human and mouse GRM1 genes on the genomic structures obtained by Crepaldi and coll. (2007). The genomic structure of the human GRM5 was established by Corti and coll. (2003). We established the genomic structure of mouse Grm5 on the basis of homology with the human gene and according to the genomic structure of it. We kept only the inter-exon sequences which are very long, to an average length of about 400000 bp. The selected words resulting from the dictionary inspection show some regular structured motifs, which are peculiar in alternating weak-binding (A, T) and strong-binding (C,G) bases. We rewrite these words in the strong-weak alphabet and perform a statistical analysis of the occurrence of these structured motifs. We extracted several classes of either omoAT or GC-rich over-represented oligonucleotides and studied their spatial distribution. Finally, we correlated the above results against the experimentally found and theoretically predicted biological properties of these inter-exon sequences available through genomic databases, like UCSC Genome.

Contact : marangon@di.unipi.it